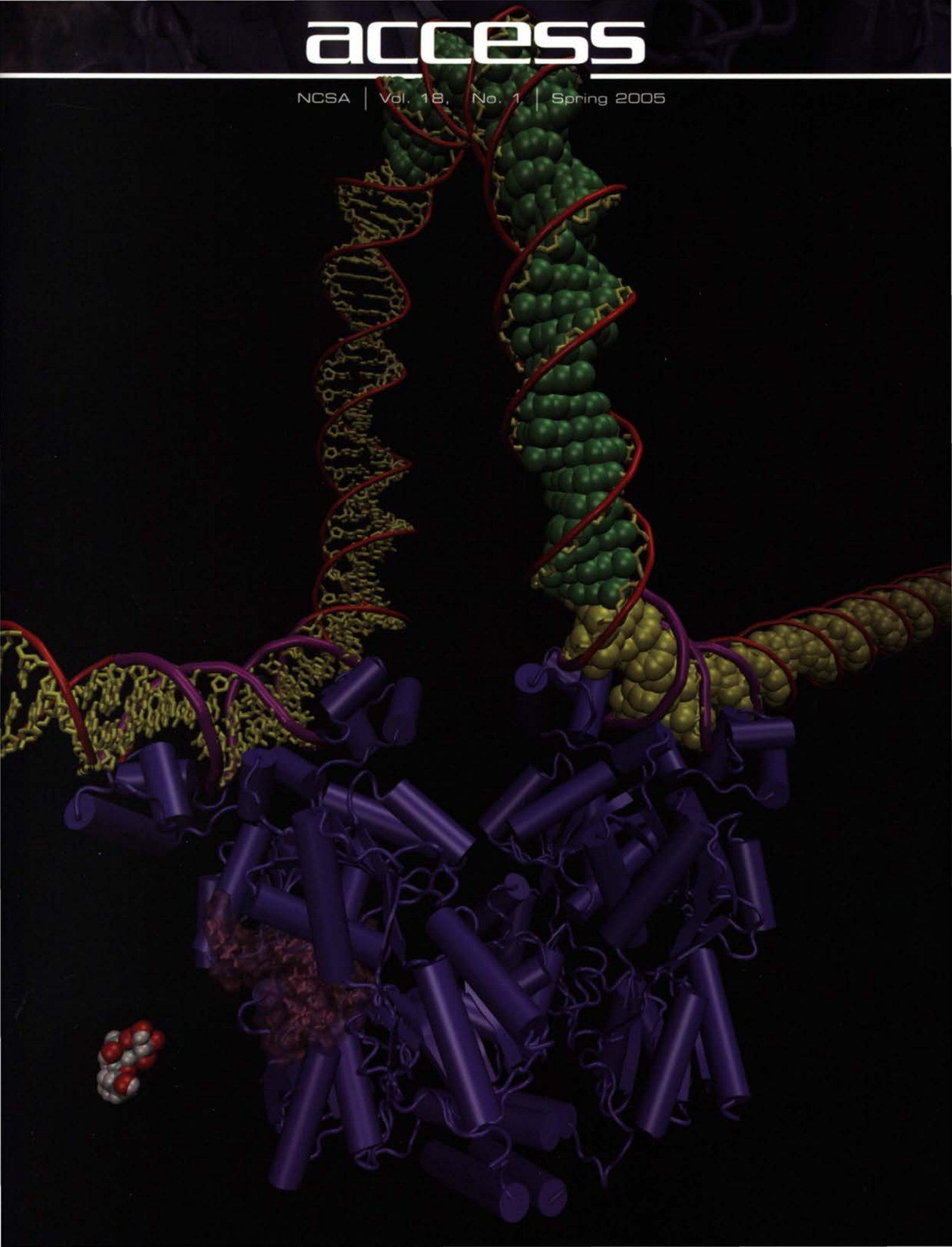
access

NCSA | Vol. 18, No. 1 | Spring 2005



access
Vol. 18 No. 1 Spring 2005

Executive Editor
Karen Green
kareng@ncsa.uiuc.edu

Managing Editor
J. William Bell
jbell@ncsa.uiuc.edu

Art Director
Carlton Bruett
cbruett@ncsa.uiuc.edu

Illustrator
Blake Harvey
bharvey@ncsa.uiuc.edu

Copyeditors
Trish Barker
Kathleen Ricker

On the Web
http://access.ncsa.uiuc.edu
Access is published on the Web every two weeks, covering the latest developments from NCSA and its partners.

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
152 Computing Applications Building
605 E. Springfield Ave.
Champaign, IL 61820-5518
217-244-0072

Who we are

The National Center for Supercomputing Applications opened its doors in January 1986. NCSA earned and maintains an international reputation in high-performance computing, networking, storage, and data mining. It is the recognized leader in developing innovative systems and software for science and engineering.

NCSA's overriding mission is to partner with diverse research communities to create the cyberinfrastructure that makes possible new scientific discoveries. Its specialty is shaping the most cutting-edge computers into working systems complete with software applications and tools for visualization, data mining and analysis, and collaboration. These innovative systems are the heart of an emerging cyberinfrastructure that links disparate systems into a single, seamless resource.

NCSA is a key partner in the National Science Foundation's TeraGrid project, a $100-million effort to offer researchers remote access to some of fastest unclassified supercomputers as well as an unparalleled array of visualization tools, application software, sensors and instruments, and mass storage devices. NCSA also leads the effort to develop a secure national cyberinfrastructure through the National Center for Advanced Secure Systems Research, a project funded by the Office of Naval Research.

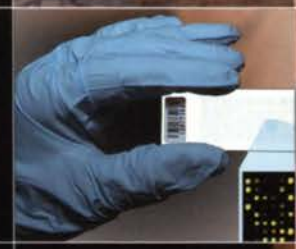
The center leaves its mark through the development of networking, visualization, storage, data management, data mining, and collaboration software as well. The prime example of this influence is NCSA Mosaic, which was the first graphical Web browser widely available to the general public. NCSA visualizations, meanwhile, have been a part of productions by the likes of PBS's *NOVA* and the Discovery Channel.

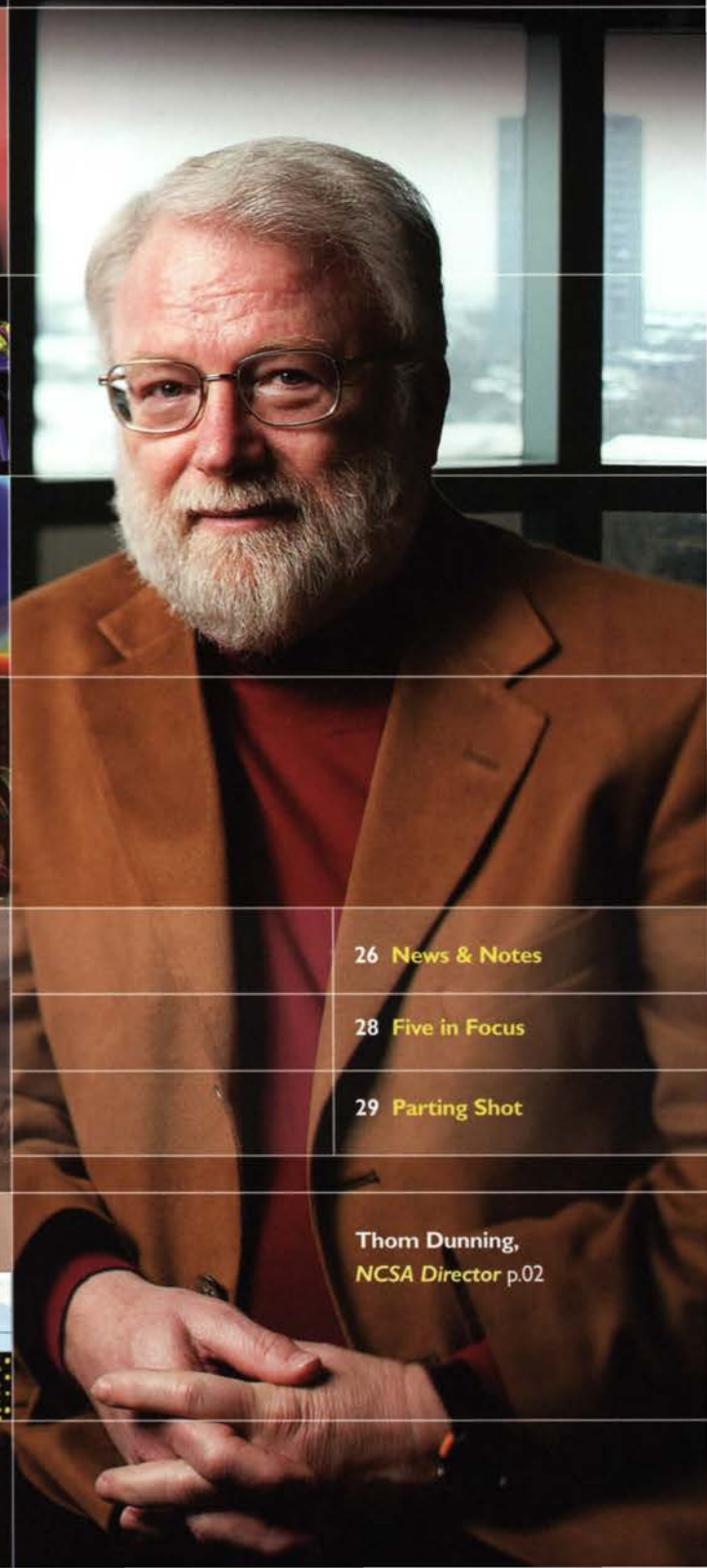
Major support for NCSA is provided by the National Science Foundation. Additional funding comes from the state of Illinois, industrial partners, and other federal agencies.

Cover
Lac repressor protein (in blue) bound with DNA (yellow and red). This simulation was created by the Theoretical and Computational Biophysics Group at the University of Illinois at Urbana-Champaign using NCSA's Mercury Linux cluster, the largest computational resource of the National Science Foundation's TeraGrid cyberinfrastructure.
See story on page 6.

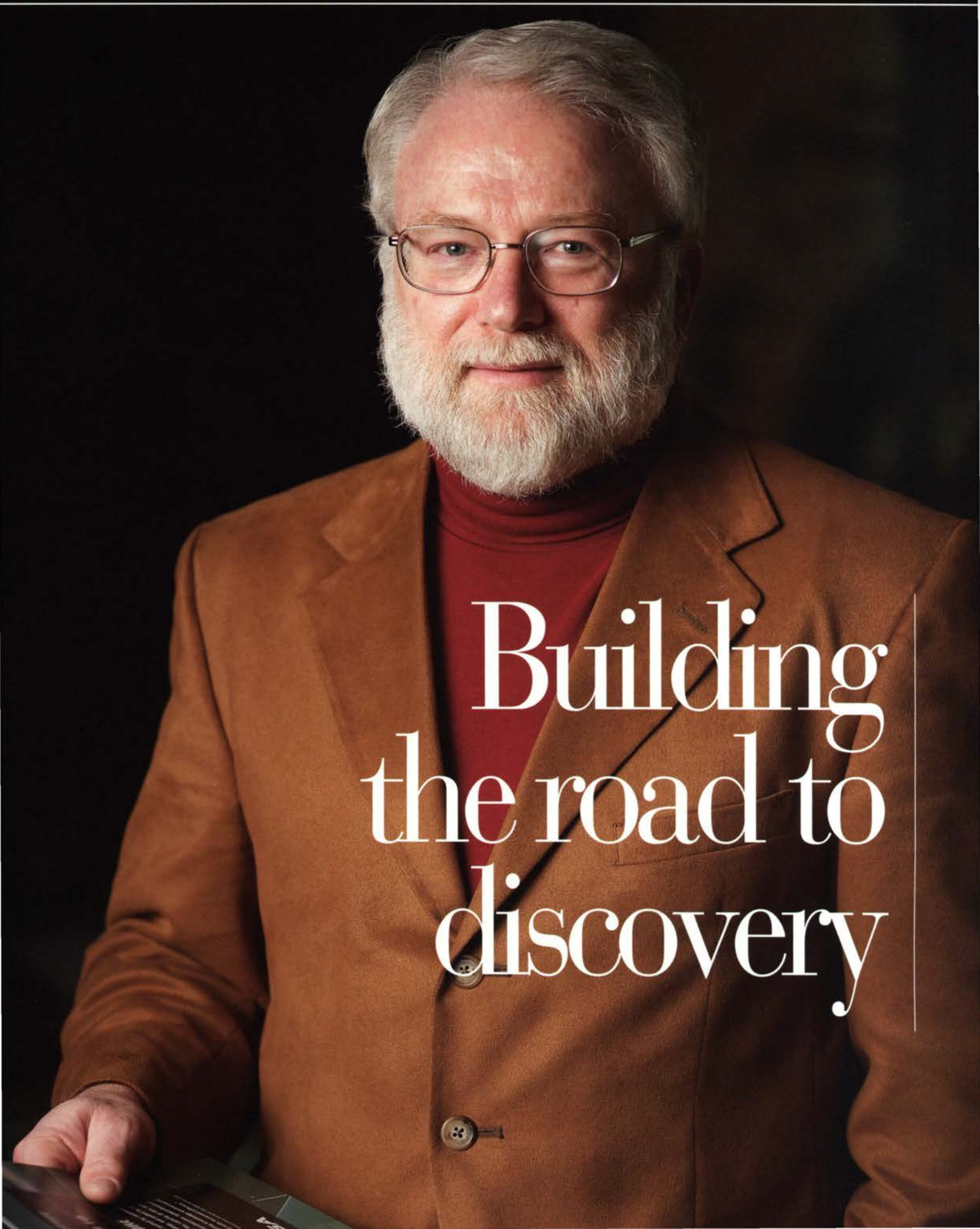
NCSA

contents





Thom Dunning,
NCSA Director p.02



Building the road to discovery

The Director's View

The development of cyberinfrastructure to advance science and engineering has reached a critical juncture. In the past, our efforts were geared toward those who needed the most powerful computing capabilities available. While many still need big iron computing power, new communities have entered the picture, and their requirements vary. Some seek to manipulate and analyze increasingly larger volumes of data located at multiple sites; others want for storage capacity. Today, we must satisfy a multitude of needs from multiple communities, and that means integrating a very diverse set of capabilities into a robust, easy-to-use cyberinfrastructure.

Engineers wouldn't build a highway without first understanding where people live and their day-to-day travel patterns. The builders would visit communities, study the lay of the land, and gather information from politicians, business leaders, educators, and soccer moms. Likewise cyberinfrastructure—that grand vision of putting the best computers, data stores, software, and research tools into the hands of scientists through an easy-to-use, secure, networked environment—must be more than the latest dazzling creation of the world's brightest computer scientists, network specialists, and software engineers. Of foremost importance are the needs of the customers, the nation's scientists and engineers. Building a national cyberinfrastructure to empower research without close collaboration with the nation's researchers is like building a highway system without understanding where people want to go. You might end up with an eight-lane road to nowhere that is used by no one.

As a scientist with more than 30 years of experience in chemistry and chemical engineering, I am one of those cyberinfrastructure customers. I know the frustration of porting a code to an unfamiliar computing system, keeping track of computations and their output scattered across various machines, and reformatting data for use in another application. I've spent many hours on what is essentially scientific busywork, whether it is transferring data or figuring out how to change the parameters of a particular computational run. I'm willing to bet that most users of high-performance systems want what I want from cyberinfrastructure: a supportive system that empowers rather than frustrates, that is seamless and as intuitive as the Web, that handles the busywork so that they can concentrate on the science.

As the new director of NCSA, I am now in charge of an outstanding team of cyberinfrastructure builders—individuals with years of experience in high-performance computing, data analysis, visualization, and networking who are dedicated to providing research communities with the capabilities they need to make the next round of discoveries. However, we cannot do our job alone. Unlike building a highway, which is a well-understood task, how to best develop cyberinfrastructure is still a topic of active discussion. And of course, there are technical challenges to overcome.

There is no other way to build this highway to discovery than to engage the research communities who are eager to take part in defining and planning cyberinfrastructure. We pledge to roll up our sleeves and work with them to bring their visions into being. We recognize that this will not be a simple sequential process of defining requirements, building the infrastructure, and testing its features. There are too many unknowns. Instead, we will launch a dialogue between the researchers who have work to do and the NCSA technologists who aim to give them the tools to do their work. The process will be labor intensive, at least at the start, which means we will not be able to work with all communities at once. However, we will work with as many as possible, and science and engineering as a whole will benefit from much of our labor.

NCSA has no intention of building a road to nowhere. Instead, we will work with scientists and engineers to build a cyberinfrastructure that takes us to places we can barely imagine now.

Thom Dunning
NCSA Director

The importance of being impatient

Shirley M. Malcom is well known as a scientist and an advocate for better science education and diversity in science and engineering. As head of the Directorate for Education and Human Resources Programs for the American Association for the Advancement of Science (AAAS), she leads programs aimed at improving science education, increasing public understanding of science and technology, and increasing participation by underrepresented groups in all areas of science. She holds a PhD in ecology and served on the National Science Board from 1994 to 1998 and on the President's Committee of Advisors on Science and Technology from 1994 to 2001. Last fall, she addressed attendees at the Grace Hopper Celebration of Women in Computing Conference in Chicago. NCSA's Karen Green caught up with her just after her talk.

Q: In your talk today, you said "There are some issues we can be patient about, and there are others that we need to be impatient about." What are the issues about which scientists and engineers should be impatient?

A: I think we should be impatient about enforcement. We have laws on the books, and yet people are still able to do job searches without really looking. They do sorts, not searches. They sort through the applications they receive. They don't actually search out the good people. I think we need to be impatient with regards to nontraditional women and minorities and the opportunities they have. For example, there are many people, often people who already have degrees, who move into the field or want to move into the field. We need to find more routes that allow them to do this. We need to see that there are opportunities to engage their skills. These are often highly motivated people with a lot to offer to our fields. That's what I mean by being impatient. I can wait to grow a new generation of children, to grow them in the right way. I have to wait. I have no choice. But sometimes you have a choice. There are already a lot of women in computer science and engineering. What are their opportunities to take on leadership roles? We have to be impatient about that.

Questions&Answers

Q: Women are well represented and have been able to excel in some scientific fields, such as biology and some of the medical fields. Why have these communities been so successful in attracting women and not computer science and engineering?

A: In some fields it's a lot more clear what the relevance is to the things people care about. I think women have been able to connect the biological sciences, for example, to the things that are important to them. In computer science and IT, the marketing has not been done in such a way that people are able to look at the field and the work being done and connect it back to the things they care about. That is the issue for IT and engineering. I don't think it has really been dealt with yet.

Q: It's been said that when women are well represented in a field, salaries on the average tend to be lower. Why?

A: In many cases we are just happy to be there, and in many cases we will accept somewhat less than what men would demand. That allows people to devalue what it is that we're bringing to the table. We're willing to be complicit when people are offering us less because we feel that the work is valuable, even if we ourselves are valued less than a colleague.

Q: None of the top universities in engineering and computer science are among those that grant the most degrees to women and minorities. Why do you think that is?

A: They can change that, they really can. But do they want to make the effort? If you can fill up your classes and fill up your program without putting forth the effort to reach out to women and minority students, why bother? If you don't think there is some additional value that diversity brings to your program, why bother? Maybe when it's hard to get foreign students to this country, there will be a lot more attention paid to the people who are already here. But it hasn't happened yet. I remember one instance where a faculty member asked for money to make a recruitment trip to China. He didn't request money for a swing through the southeastern United States. A question that needs to be asked is what are [the top schools] looking for in the students they recruit? Maybe they are asking students to have a record of achievement in an area that hasn't even been made available to them. To be very candid, there are some departments that make me wonder what would have happened if they had looked at my record and judged me only on my record. If that were the case, I wouldn't be here today. They would have to bring me in on my potential. So it's a question of potential and capacity, not necessarily just test scores.

Q: Data show that the American work force will be primarily nonwhite and include more women than men by 2020. What are the implications of these demographics to the IT field?

Q: Shirley Ann Jackson, who is the president of AAAS, calls this the underrepresented majority. When you look at women, who are about half the workforce, and you add to that African American, Hispanic, and Native American males, that's a majority of the people. If you ignore the majority, how can you expect to be successful? For one thing, I think that the products that the IT industry makes could very well be out of phase with their markets. That's where we are. There hasn't been a real understanding of that or an acknowledgement of that.

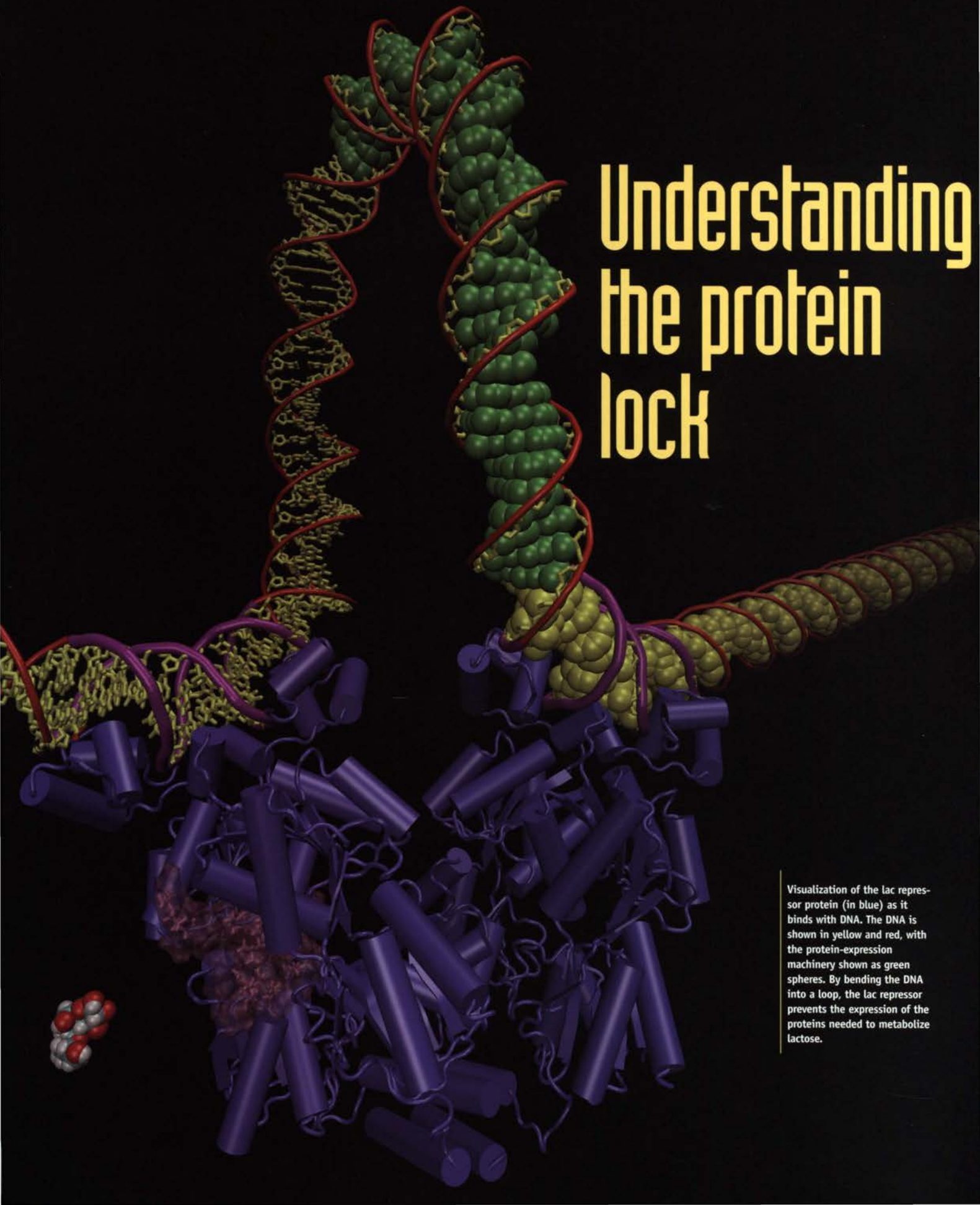
Q: The IT culture is known for demanding long hours. Is that culture changing at all to accommodate the needs of people who may want more balance in their lives?

A: I honestly don't know if that culture is changing or not. If you want to talk long hours, I can put my own life up against anybody's long hours. I'm on the road a lot, I have a lot of demands, and I spend the time I need to spend to get things done. And I have children. They're adults now, but that was part of it too. So it's not so much a question of long hours but a question of whether that culture allows variation in the way that you manage to get those work hours in. For example, I worked a regular work day when I could, then I went home, I spent time with my children, I supervised homework, cooked dinner, and after that I'd start again with what needed to be done for work. Does that startup company, that IT company, require that they see your face all the time or will they be satisfied that the work is getting done even if it is getting done remotely? One of the things about technology is that because it allows you to access things remotely and work remotely, your presence or absence in a space doesn't mean much, except in the context of having to demonstrate to an administrator that you are working. One of the things I tell my staff is that showing up is not what I require, it's getting the work done. People have different styles of work, and I think there has to be room for different styles of work.

Q: Considering how rare women, particularly African American women, are in computer science and engineering, do you think they can truly be themselves in the workplace? Or do they have to change in order to thrive?

A: I think they can be themselves. I think that you have to hold on to who you are with both hands because there often is this pressure to homogenize. But that's not why you were hired. That's not what you bring to the table, that's not your value to the organization. Your value is that you are different and you don't just blend in with the scenery. One of the things to look out for and to avoid is the expectation that you will bring only that different perspective and nothing else. When I was a member of the National Science Board, I was not going to let anybody tag me as just "the diversity person." I was there to deal with the issues, whether they were high-performance computing, long-term ecological resources, or whatever. If the question was about engineering research centers, I was there to ask about engineering research centers. I might also ask about diversity, but it certainly wasn't the only thing.

Understanding the protein lock



Visualization of the lac repressor protein (in blue) as it binds with DNA. The DNA is shown in yellow and red, with the protein-expression machinery shown as green spheres. By bending the DNA into a loop, the lac repressor prevents the expression of the proteins needed to metabolize lactose.

by Trish Barker

Using a novel multiscale approach, researchers at the University of Illinois gain insight into a mechanism that suppresses gene expression.

An organism's genome contains all of the information required to build everything—all of the organs, cells, and cellular structures—the organism will ever need. The trick is building the right thing at the right time.

At any given time, DNA needs to be activated—to express a key protein, for example—or suppressed, holding back the potential to express that protein until it is needed.

Klaus Schulten, leader of the Theoretical and Computational Biophysics Group at the University of Illinois at Urbana-Champaign, wanted to explore the mechanism that controls when a gene triggers expression of a protein and when that expression is held back. To tackle that question, his group used a novel multiscale approach that required both mathematical modeling and computational simulation using NCSA's Mercury Linux cluster, the largest computational resource of the National Science Foundation's TeraGrid cyberinfrastructure.

Lactose metabolism in *E. coli*

Schulten's group decided to examine a prototypical case of gene suppression in the bacterium *E. coli*.

When lactose is available in the environment, the bacterium needs to import and metabolize it. A set of genes, known as the lac operon, encodes the three proteins needed for this process. When environmental lactose is not available, the proteins are not needed, and the lac operon needs to be suppressed.

A protein known as lac repressor is responsible for holding these genes back. The lac repressor finds the lac operon segments of the genome, bending this stretch of DNA into a loop. This loop prevents expression of the lactose-metabolizing proteins.

Lactose is the key to this DNA lock. When the sugar becomes available in the bacterium's environment, it diffuses into the cell and unlocks the lac repressor. The loop is unleashed and the genes are expressed, resulting in the production of the three proteins that are necessary for lactose metabolism.

"This protein has an exemplary role in controlling genetic information," Schulten explains, which makes it an ideal candidate for research.

Beyond crystal forms

While the broad outlines of the loop mechanism employed by the lac repressor were understood, Schulten wanted to look at the system in more detail. Knowing the structure of the lac repressor protein would enable the development of a detailed description of how it functions—how does the lac repressor grab DNA, bend it into a loop, and hold the loop in place despite its resistance?

The typical way to determine the structure of a protein is to crystallize it into a rigid, repetitive form and then examine the geometry of the crystal. But in this case, the protein in action is moving and flexible. Attempts to crystallize the protein with the DNA loop in place failed, showing just the protein and the binding sites on the DNA strand, not the actual loop.

"[We] needed to describe the dynamics of the protein," Schulten says. In fact, the paper in which Schulten and two members of his research group, postdoc Alexander Balaeff and graduate student Elizabeth Villa, have described their lac repressor research is titled "What the Crystal Did Not Show." The paper is appearing this spring in the *Proceedings of the National Academy of Sciences*.

A multiscale approach

Capturing the form of a protein in action is a daunting and computationally intensive task to tackle with molecular dynamics (MD) simulation. The lac repressor is a huge protein; modeling its behavior in a realistic water/salt environment leads to a simulation size of more than 200,000 atoms. Adding the DNA loop increases the simulation to at least 700,000 atoms, a size that could not be sustained for sufficient time even using today's most powerful supercomputers.

To reduce the computational cost of the simulation, Schulten and his group devised a dual, or "multiscale," approach that combines mathematics and computing.

Visualization of the multiscale simulation. The structure of the complex formed by the bound lac repressor (shown in green) and DNA is simulated inside an ion-filled water box. The DNA loop connecting the protein-bound DNA segments is modeled as an elastic ribbon (shown in gray); the forces of interaction between the loop and the protein-bound DNA segments (red arrows) are included in the molecular dynamics simulation.

Using the NAMD parallel molecular dynamics code developed by Schulten's group for high-performance simulation of large biomolecular systems, the researchers simulated the motion of the lac repressor protein in a water/salt environment. The motion of the DNA loop was modeled by solving the system of Kirchhoff equations of elasticity, a complex set of differential equations.

"The mathematical description of DNA and the computational modeling of the protein had to communicate with each other," Schulten says. The mathematical model of the DNA loop provided the MD simulation with the forces with which the DNA tried to resist the formation of the loop, while the simulation provided the model with information on how the ends of the loop moved.

Molecular dynamics simulations were carried out at both the Pittsburgh Supercomputer Center (PSC) and NCSA. At PSC, a system of more than 200,000 atoms was simulated for 22.4 nanoseconds using 600 processors; the average production speed was 2.7 nanoseconds per day. The researchers then used the more powerful Mercury cluster at NCSA to simulate a system of more than 300,000 atoms that included a more expansive water/salt environment and allowed them to observe larger conformational changes. On Mercury, 254 processors were used to simulate the system for 17 nanoseconds. The production speed was 2.5 nanoseconds per day, nearly equal to the speed achieved at PSC where more than two times the number of processors were required.

Schulten says the group's simulations on the TeraGrid system at NCSA represent one of the first cases of using a multiscale approach for the description of an important biological system, explaining that this approach required the availability of a computational system that could handle the simulation of such a large protein system.

"We took advantage of the computing power at NCSA," he says.

What the crystal did not show

When the researchers examined the results of their multiscale approach, they could show for the first time the entire lac repressor/DNA complex, what Schulten calls "one of the most magnificent biopolymer molecules."

"[We] saw how the protein wrestles with the DNA," he explains. They saw "the ingenious way by which the protein overcomes the resisting DNA—with extreme flexibility and, for lack of a better word, patience."

They learned that it is not through sheer size that the lac repressor protein overpowers and subdues the DNA. It was previously known that the protein has two arms with which to grasp the DNA binding sites; Schulten's simulations showed that these arms grasp the DNA with heads that connect to the protein with thin coils. These coils provide the protein with extreme flexibility. "No matter how the DNA tries to wriggle and bend," Schulten says, "the protein can follow and maintain contact."

In fact, the researchers simulated the application of exaggerated forces to the protein heads. Even under forces 50 times stronger than a real-world system, they found that the lac repressor could maintain its shape and its grip on the DNA loop.

The multiscale combination of mathematical modeling and computational simulation employed by Schulten's group not only provided insight into the structural dynamics of a particular biomolecular system, it also provided a pathway for future investigation. Similar protein-DNA complexes form in the genomes of many living organisms, so the multiscale approach could lead to advances in our knowledge of living things in general, of the human body, and of medicine.

"The story of the lac repressor/DNA loop simulation on the TeraGrid's Mercury system is an exemplary case for computational science, showing how the combination of modern computational techniques, of wise investment, of hard work and great ingenuity leads researchers to learn what experiment cannot tell us," Schulten says. "The computer is becoming more and more like a new microscope that permits views into the world that cannot be obtained by other means."

This research is supported by the National Institutes of Health.

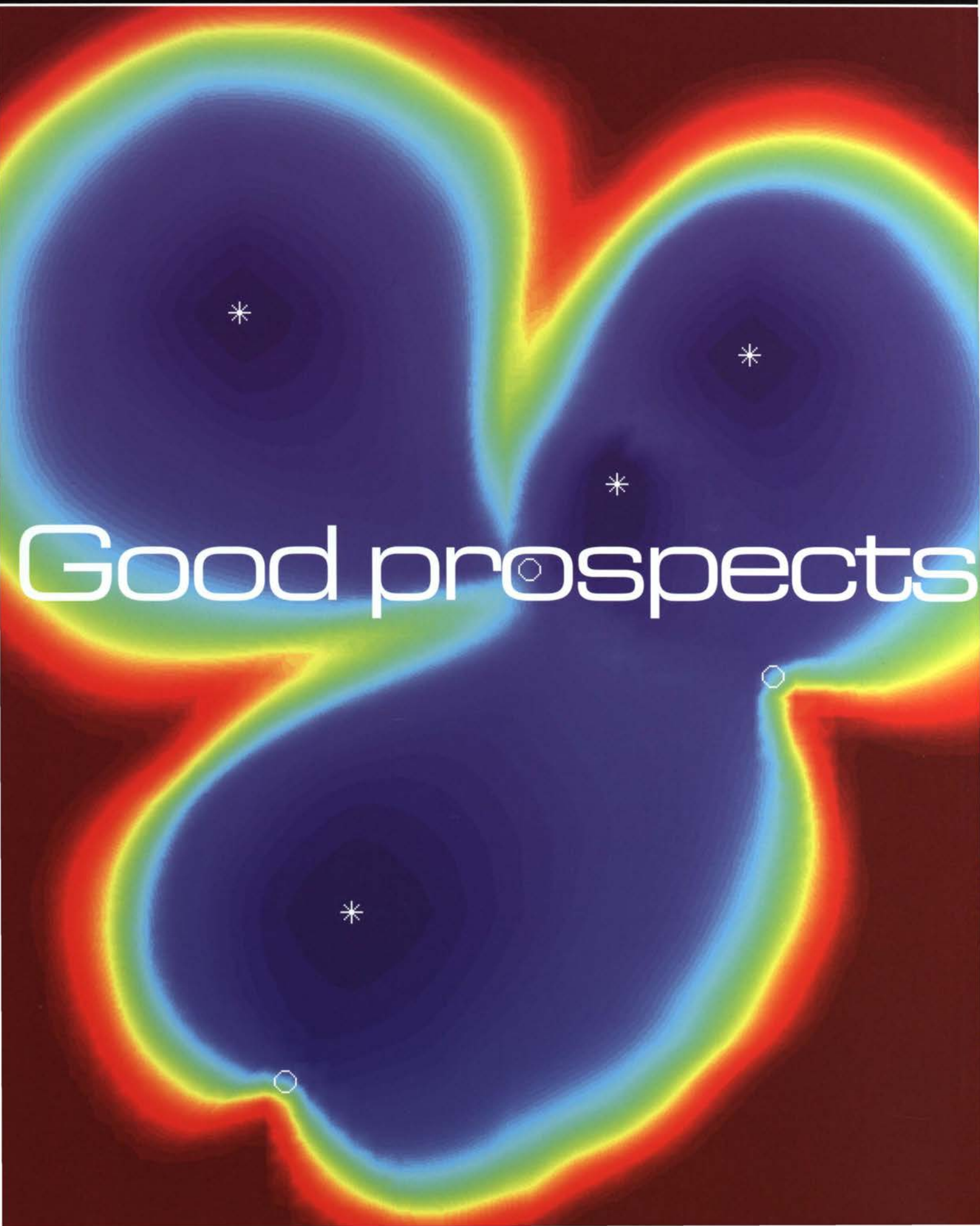
Access Online: http://access.ncsa.uiuc.edu/CoverStories/lac/

For further information: http://www.ks.uiuc.edu/

Team members

Alexander Balaeff
Klaus Schulten
Elizabeth Villa

Lac repressor protein (shown in red) wrestling with DNA. Various structures (shown as colored ribbons, drawn every 100 picoseconds) develop during the simulation. The lac repressor holds its shape during this process, with only the flexible head groups rotating to combat the strain.



Good prospects

by J. William Bell

Seismic modeling and reservoir simulations come to the TeraGrid, improving two workhorses of the oil industry.

Oil prospecting used to rely on hunches as much as anything else. Where to explore? Luck might land you a spot where oil was seeping from the ground. Where to set up the wells, pumps, and other equipment? There were some rules of thumb to fall back on. How to manage production? Perhaps you'd check the site's current performance, compare it to past wells' behavior, and then go by instinct.

Nowadays, however, a golden gut isn't worth what it used to be.

"Twenty or 30 years ago, people started to wonder how to get more from their wells," says Wolfgang Bangerth, a postdoc at the University of Texas at Austin's Center for Subsurface Modeling. "They began to wonder, 'What are the clever ways to do this?'"

Some of these more clever ways are newer techniques in drilling and production. Companies can now drill horizontally and sink wells that branch in multiple directions. They can shut down sections of their wells. Pumps, which force oil toward wells, can be run at different rates and on varying schedules.

Companies also need to choose ideal places for this equipment and to surmise the geological features of the ground beneath it. That's where people like Bangerth and his collaborators come in. Using TeraGrid resources, a multidisciplinary team from UT, The Ohio State University, and Rutgers University is at work on software tools that improve companies' oil reservoir management. The tools allow them to better exploit existing reservoirs, find new reservoirs, and minimize drilling's adverse environmental impact.

"Reservoir models, which help guide production strategies, and seismic imaging, used to determine subsurface properties, are two workhorses of the oil industry," Bangerth says.

A reservoir of knowledge

Reservoir models are quite complicated. They account for different fluids like water, oil, and gas, different rock properties, the features of pumps and production wells, and sometimes complicated chemical reactions. To make modeling these complex systems computationally tractable, mathematicians subdivide the reservoir into a mesh of blocks. They then associate wells, pumps, and other equipment with individual blocks and solve an approximate model of the systems' fluid dynamics. The team uses IPARS, a multi-model, multi-phase reservoir simulator developed at the Center for Subsurface Modeling under the direction of Mary Wheeler. The output of IPARS is translated into production rates and ever-important revenue levels. Equipment is then moved around the mesh in order to compare different configurations and to find the best one.

The possibilities aren't endless, but their sheer number sometimes makes researchers pine for the days of hunch-based prospecting.

"You can have billions of possible configurations that need to be examined, so you can't just do an exhaustive search of the parameter space [the collection of all possible configurations within a given grid]," according to Tahsin Kurc, an assistant professor at Ohio State and part of the Multiscale Computing Lab that is led by Joel Saltz. A single IPARS run usually takes hours. If it's really difficult, hours can bleed into days.

"Complexity usually translates into precision," Kurc says. "We want to move intelligently through the search space."

Intelligent movement relies on a dynamic, data-driven optimization system. Large volumes of data obtained from earlier simulations and dynamically updated by new simulations or experimental measurements are stored, queried, and analyzed to find promising initial configurations. These configurations are then refined with on-the-fly monitoring and steering of the simulation and optimization processes.

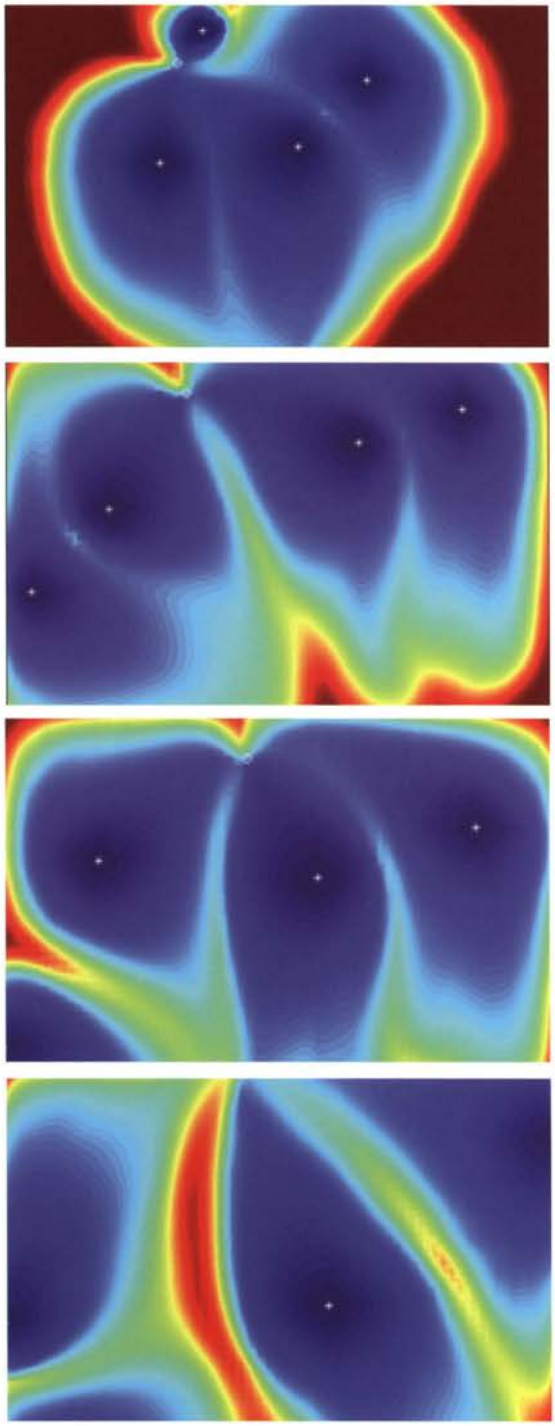
A set of simulations provides a rough sampling of the search space. Middleware tools from Saltz's team, called STORM and DataCutter, manage the very large amounts of data produced by these simulations. These tools are also used to identify good starting points for more comprehensive searches. Dynamic steering and collaboration tools—AutoMate and DISCOVER from Associate Professor Manish Parashar's lab at Rutgers—allow on-the-fly searches within these subsections. Sophisticated optimization algorithms coupled with IPARS models guide these searches by comparing configurations in the subsections.

'Do the math on that'

High oil output isn't always the objective. Sometimes companies want to modulate output over time to allow for changes in markets and prices. Sometimes they want to find a design that comes with the lowest risk of failures or mishaps. Regardless of what result they are looking for, they want to identify that result in the shortest possible amount of time. Distributed computing is key.

"We're supporting this now in a grid environment, modeling multiple configurations and multiple points concurrently," Kurc says.

The team recently completed a set of about 25,000 runs—each taking about two hours on a single processor—in less than a week. "Do the math on that," Bangerth says. "You're talking 200 to 400 runs going at any one time...That's not something we're used to having."



Visualizations of oil reservoir simulations at various stages in an optimization. Pumps, represented by asterisks, push oil toward wells, represented by small white circles, that draw oil from the ground. Blue areas indicate areas of high water concentration. Brown areas indicate areas of high oil concentration. By the end of the simulations, there is not much oil left in the reservoir, and what is left has been swept toward the wells.

These calculations were completed using the TeraGrid cluster at NCSA. More runs are ongoing at NCSA and other TeraGrid sites across the country, including the San Diego Supercomputer Center and the Texas Advanced Computing Center. Machines at UT's Institute for Computational Engineering and Sciences, host institution for the Center for Subsurface Modeling, are also in use. The TeraGrid is the world's largest, most comprehensive infrastructure for open scientific research. It includes 20 teraflops of computing power distributed at nine sites, facilities capable of managing and storing nearly one petabyte of data, high-resolution visualization environments, and toolkits for grid computing.

NCSA's Bruce Loftis and Byoung-Do Kim helped port the reservoir simulation and optimization codes to the TeraGrid machines and built a toolkit that simplifies execution across multiple systems. The toolkit makes it easier for researchers to "babysit," as Kurc calls it, these large, distributed runs. It shows where calculations are taking place, which are complete, which have failed, and which are ongoing. That's no small feat when orchestrating work across hundreds of processors.

"The TeraGrid is well suited to this project—massive numbers of independent jobs," Loftis explains. "There are lots of these sorts of problems out there."

At the refinery

Oil reservoirs are generally inaccessible, being thousands of feet under the ground or the ocean. Unsurprisingly, little is typically known about their exact geological features. Scientists

cope with this problem in two ways. Either they solve their equations on hundreds of geological models that are equally compatible with the existing knowledge of a site, or they try to come up with additional information.

Some information on geological conditions can come from real-world tests. Sound waves are blasted through the earth and bounce back to receivers on the ground or the ocean's surface. This echo can be translated into information about things like the type, density, and permeability of the rock and the amount of oil contained within. These soundings are taken over and over from different positions for a single area. They're expensive propositions, especially if they are to capture the level of precision researchers really want for their optimization.

The team's seismic models are used to develop likely geological conditions, based on simulated soundings. These conditions fine tune the reservoir models, making them as realistic as possible from the get-go.

The results of a single sounding passing into the ground and bouncing back can consume 20 gigabytes of space. Currently, more than eight terabytes of seismic simulation data sit on clusters at NCSA, ready to be integrated into the reservoir models. With the distributed storage and computing power of the TeraGrid and the middleware tools STORM and DataCutter, the team is looking to create more than 10 times that amount in the short term.

"And a good, big [seismic survey of an area] would be into the petabytes," Kurc says.

The team hopes to end up with a system that allows those prospecting for oil to build a database of possible conditions that have already had reservoir optimizations run. Companies will assess the geological features of the site they are interested in, query the database for the description that most closely resembles the site, and receive an already-completed optimization in return.

It might take some of the romance out of the process when compared to the make-or-break old days. But what's a little romance in the face of fewer failures, fewer environmental problems, and more dollars?

This research is supported by the National Science Foundation's Information Technology Research program, the Department of Energy, and the Department of Defense.

Access Online: http://access.ncsa.uiuc.edu/CoverStories/oil/

For further information: http://www.ices.utexas.edu/csm/
http://multiscalecomputing.org/
http://automate.rutgers.edu/

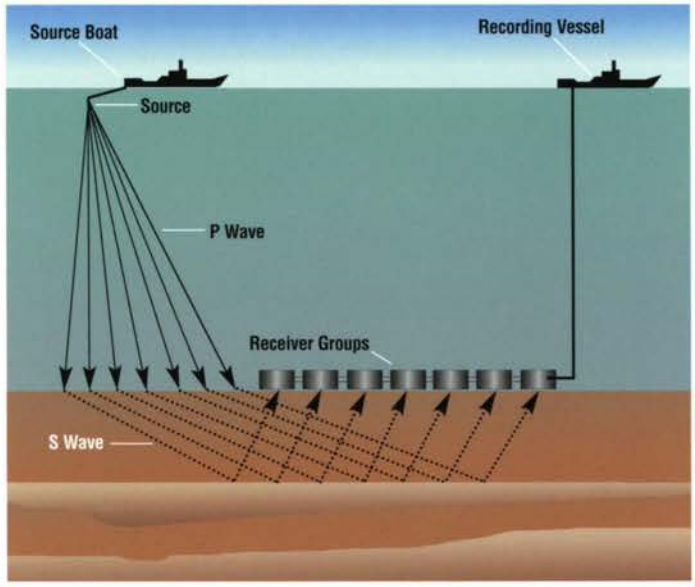
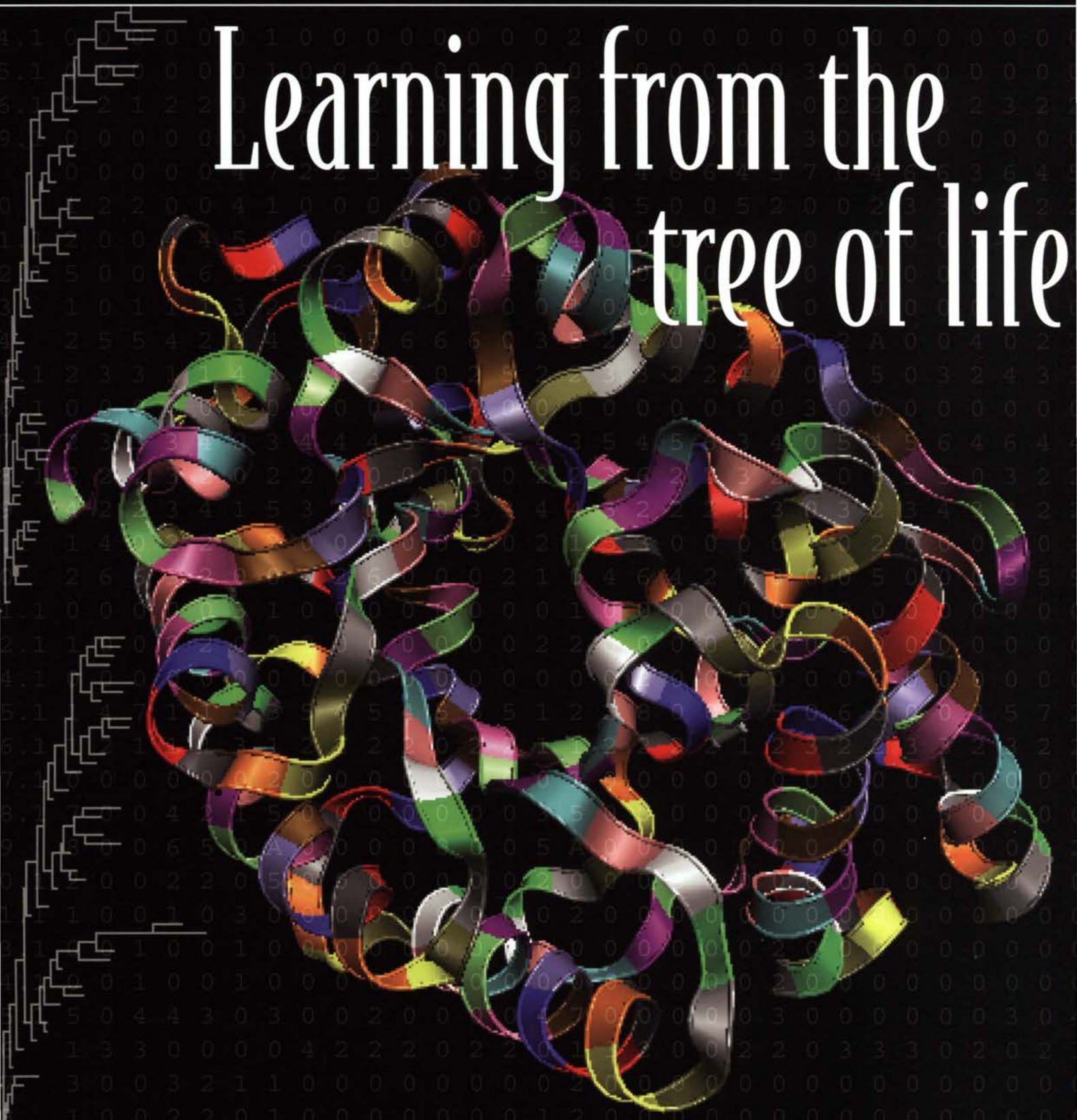


Illustration of soundings being taken on the ocean. Sound waves are projected from a ship, pass through the water, and are altered by the rock in the ocean floor. The echo from this process is picked up by receivers on the ocean floor or seismic monitoring platforms. It can be translated into information about things like the type, density, and permeability of the rock and the amount of oil contained within. These soundings are taken over and over from different positions for a single area.

Team members

Wolfgang Bangerth
Viraj Bhat
Umit Catalyurek
Shannon Hastings
Hector Klie
Vijay S. Kumar
Tahsin Kurc
Steve Langella
Vincent Matossian
Scott Oster
Manish Parashar
Benjamin Rutt
Joel Saltz
Roustam Seifoullaev
Mrinal Sen
Krishnan Sivaramakrishnan
Paul Stoffa
Mary Wheeler
Michael Zhang

Learning from the tree of life



by Kathleen Ricker

Funded by TRECC, a UIUC biochemist explores a possible link between the tiniest protein molecules and some of the biggest events in human history.

What do protein molecules—some of the most fundamental components of life—have in common with the paper clip, the internal combustion engine, or a successful blue-chip stock?

Evolutionary fitness, argues Gustavo Caetano-Anolles, a biochemist at the University of Illinois at Urbana-Champaign. "At some point, someone invented the paper clip. There's a rationale for why the current design is the one we use now, rather than some other. There's a rationale why we have a car and not something different. There's a history linked to this, an inherent entity that defines why a car is what it is, or why a paper clip is what it is."

History, aesthetics, practical design, even random accident all figure into the evolution of the technologies and institutions on which we rely heavily today—and in each case, fitness determined the outcome of that evolution, just as fitness to perform a particular function determined the structure of the protein molecule. "Fitness is ultimately the element that defines the success of a particular molecule, or of any organism or entity you study," says Caetano-Anolles. "If you start thinking about the different elements we humans use and have generated as inventions, they also have fitness components. Those that are not fit disappear."

A researcher in the Department of Crop Sciences at UIUC, Caetano-Anolles is examining how the evolutionary mechanisms of protein molecules may also have implications for more complex systems and organisms. Caetano-Anolles's research currently receives support from the Technology Research, Education, and Commercialization Center (TRECC), a UIUC program funded by the Office of Naval Research (ONR) and administered by NCSA. TRECC supports innovative research in advanced information technologies and their application for the Navy R&D community.

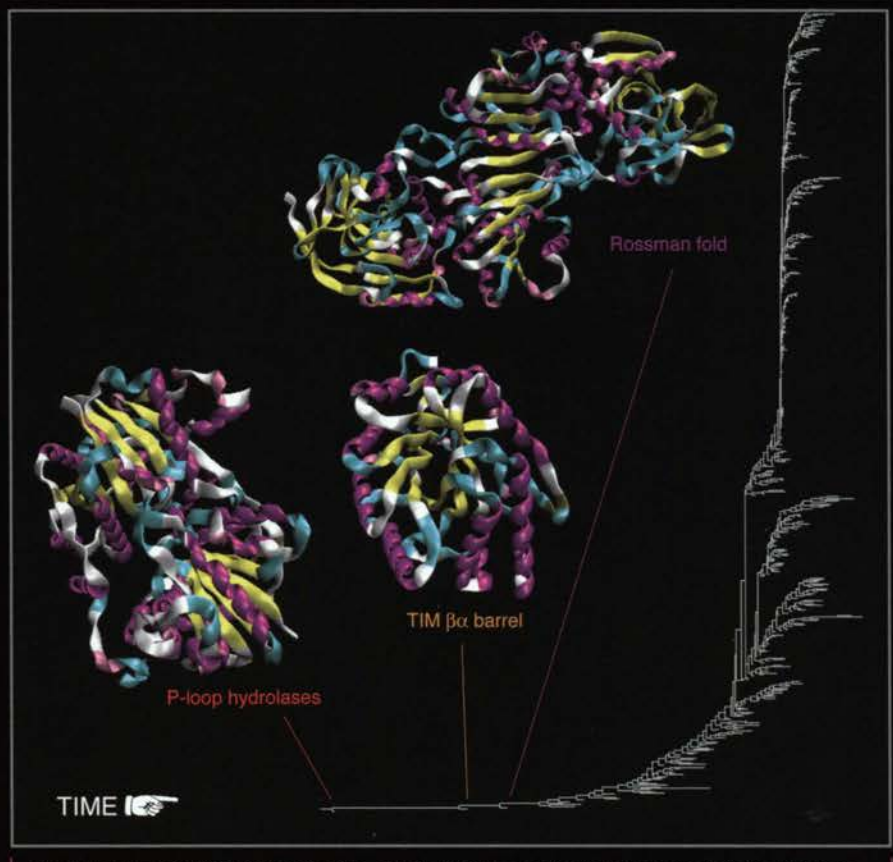
A molecular *Systemae Naturae*

Protein structures resemble bead necklaces, where the "beads" are amino acids. It is well known that how the amino acids that make up the protein necklace are arranged affects the actual structure of the protein. Furthermore, how the protein folds (what shape the amino acid chain takes when coiled upon itself) is determined by the protein's biological function. Thanks to genomics, there is now an enormous amount of data about the sequence and structure of proteins, which is now being used to generate a classification of protein architecture.

Caetano-Anolles likens this effort to that of 18th-century naturalist Carolus Linnaeus, who created a taxonomy of organisms—a "tree of life" that, over the centuries, came to extend from one-celled plants and other microscopic organisms to enormously complex, highly-advanced sentient mammals and is still undergoing expansion today. "We go to the deep vents in the ocean, and we're still isolating new bacteria, new archaea, new organisms, and groups of organisms that don't fit within our previous classifications, so we're still exploring and discovering our world," says Caetano-Anolles. "And with proteins we're doing the same thing—we're trying to understand just how complex their world is."

Proteins are classified into different groups based on their underlying architecture. The question for Caetano-Anolles, however, is what the logic behind the classification really is—and whether it is something that could be a useful model in areas other than biology. To answer this question, he is examining how proteins have evolved based on their structure. He begins not at the bottom, with the amino acids, but at the top, with simple organisms such as bacteria and eukaryotes, and works downward.

◀ **A model for the TIM __ barrel structure of a xylanase protein. Visualized using VMD software in ribbons format, it shows how the molecule 'folds' in three-dimensional space and is colored according to the different amino acids that make up the molecule. This fungal enzyme deconstructs plant cell wall materials, producing short-chain oligosaccharides and is useful for paper bleaching. Protein fold structures like these can be classified in evolutionary terms using tree representations obtained from data matrices that depict genomic demographic characteristics (both portrayed in the background).**



A universal tree of protein fold architecture, showing the three most ancient structures that rise at its base. Ancient folds share a common architecture of sheets (in yellow) and helices (in purple) that form either barrels or are interleaved and are highly symmetrical. Example proteins (from left to right) showing these structures include the nitrogenase iron protein from *Azotobacter vinelandii*, an enzyme important for nitrogen fixation, the xylanase from *Penicillium simplicissimum*, and an alcohol dehydrogenase enzyme from humans.

After classifying the proteins that appear within the genome of a given organism, he builds mathematical hierarchies, or phylogenetic trees, which show the relationships between proteins as evidenced by their architectures.

Doing so, Caetano-Anolles explains, allows him to examine how two protein folds might be related to each other—and possibly to an extinct third protein fold. Protein folds, he says, are especially useful for evolutionary research, because they have changed so little in hundreds of millions of years. "It's an ideal molecule for study," he says. "There's a rationale for that. If something is working very well, nature will try to preserve it and will keep the design static until, perhaps, some big revolution occurs, producing a new design that works better."

The mechanism of creation

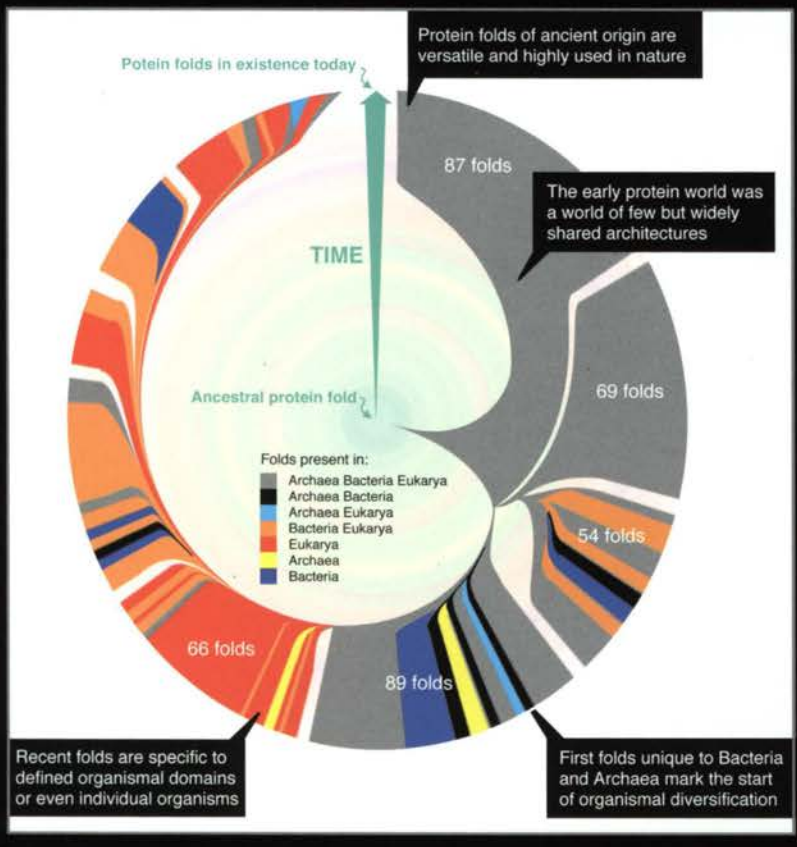
But what, exactly, is it that determines which proteins succeed and which ones have been eliminated? And how can the presence of certain proteins in a genome, and not others, help illuminate the evolutionary mechanism of a complex organism? The answer, says Caetano-Anolles, may have to do with RNA, a ribonucleic acid essential for turning the genetic information encoded in DNA into proteins. Like DNA, RNA consists of amino acids strung together but in a single, rather than a double, strand. Messenger RNA interacts with the ribosome, a kind of infinitesimal machine within the cell that that is made of ribosomal RNA and proteins and actually performs the protein synthesis. "It's really what makes life, basically," says Caetano-Anolles. "It's the central element that turns information into something that works out and produces the complexity that we see, both its structure and all the catalysts that convert chemicals or light into energy and fuel the cell's workings."

It's not hard to draw parallels between cellular structures and mechanisms and computational processes. DNA resembles a repository of stored, encoded data and the RNA-encoded ribosomal molecule are essentially the central script for a machinery that converts the data into usable output—here, the crucial proteins. The structure of the ribosomal script depends on the function it needs to perform. "If the structure you're studying changes," says Caetano-Anolles, "you can use the information to model the evolution of the molecule."

This model is useful, says Caetano-Anolles, because present-day molecules can be used to reconstruct the past and to study trends that indicate future evolutionary development.

"We're exploring whether we can devise a system that will act also as a predictive element, looking not only to the past but using what we know from the past to project into the future," he says. "Are there tools that we can use, based on the knowledge systems of biology that have been in use for a long time, that we can combine with technologies like neural networks and machine learning techniques to predict future events?"

If the answer is yes, Caetano-Anolles believes that the result could be a powerful system that combines modeling, phylogenetic analysis, and new data mining and computational analysis tools to predict outcomes in many different fields, such as engineering, social science, biology, military science, or national security.



Cartoon describing a tree diagram in circle format and some conclusions that are derived from reconstructing universal trees. At the tip of branches are protein folds in existence today; branches are pooled to describe groups of folds. At the center of the spiraling circle is the ancestral protein fold, a hypothetical primitive entity that gave rise to the protein world of today. Coloring describes how widely shared are the fold architectures between the three organismal domains, the microbial Archaea and Bacteria and the complex Eukarya (to which humans belong). The arrow indicates the flow of time.

"That's far removed from what we're doing right now, but the principle is what's important," Caetano-Anolles says. "If you look at human beings as perfect machines—I think we're more than that, of course, but it's a perfect example of how this process of change has ultimately resulted in these fantastic entities."

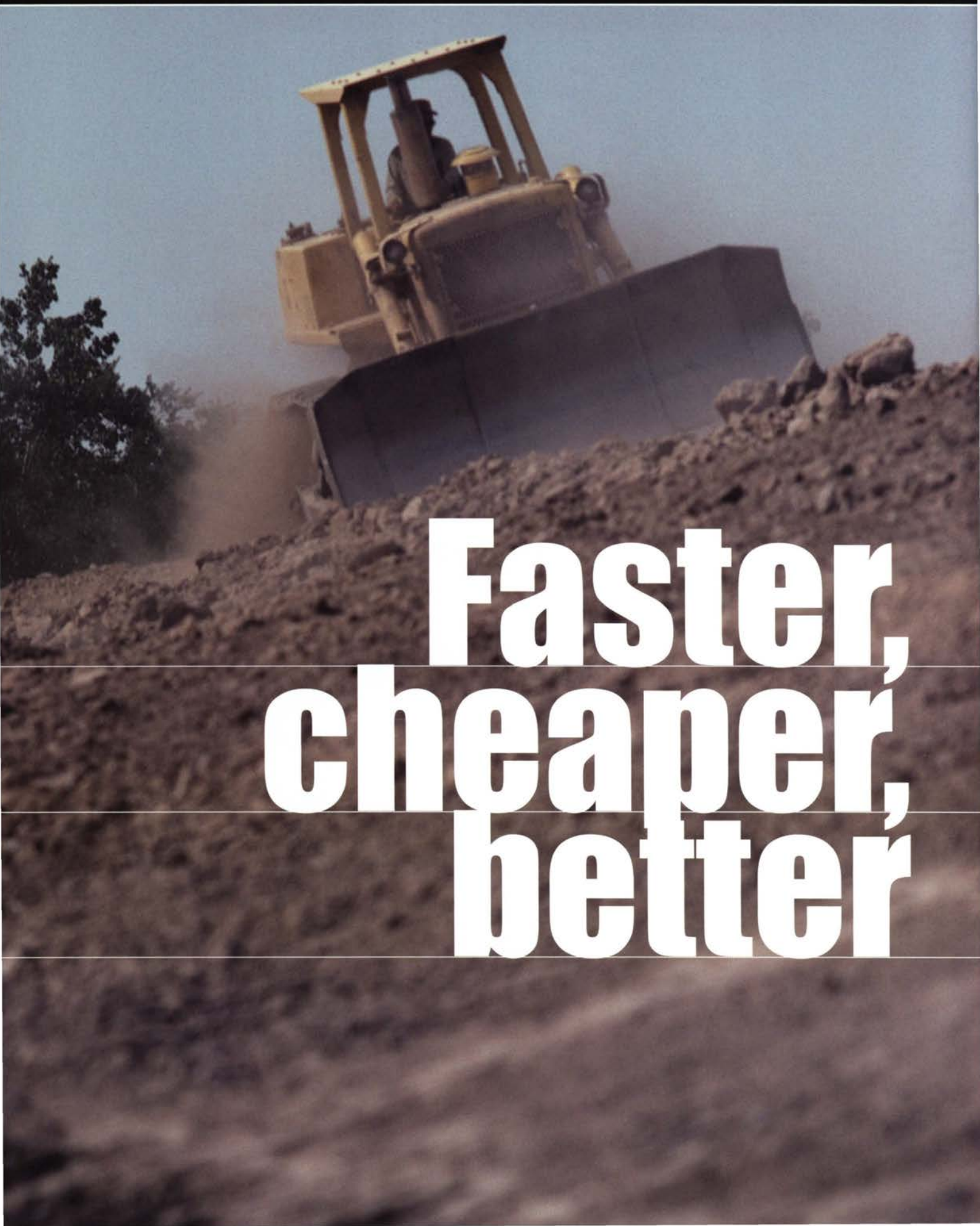
This research is supported by the Office of Naval Research and the National Science Foundation.

Access Online: http://access.ncsa.uiuc.edu/CoverStories/tree/

For further information: http://www.cropsci.uiuc.edu/faculty/gca/

Team members

Derek Caetano-Anolles
Gustavo Caetano-Anolles
Seungwoo Hwang
Hee Shin Kim
Jay Mittenthal
Feng-Jie Sun



Faster, cheaper, better

by Kathleen Ricker

Genetic algorithms could help civil engineers and planners avoid construction headaches—or, at the very least, minimize the pain.

A visualization of the time-cost-quality tradeoff problem inherent in any large-scale construction project. A reduction in the project's duration—and, hence, a reduction in the disruption of traffic patterns—comes at a higher cost, while keeping costs down is likely to result in a decrease in quality.

You're a commuter, your daily rush-hour ordeal made even more grueling by the hassle of unexpected merging lanes, the heady essence of asphalt, and the sign-toting, orange-clad road crew ahead. Resurfacing the road again? you think. But they just did that two years ago! Is this why my taxes are so high?

When it comes to major public construction projects, it's not just the public who wants the end product to be faster, cheaper, and better. The Federal Highway Authority estimates that a staggering $94 billion will be spent on transportation infrastructure every year for the next 20 years. Not surprisingly, state and federal transportation departments want to make sure that their significant infrastructure investments are worthwhile—and they've upped the stakes. The traditional bidding process, in which the least expensive estimate wins the contract, has undergone a transformation in recent years. Cost is no longer the primary factor in determining who gets the job; now project duration—the amount of time that drivers will be negotiating the construction—and quality and durability are also important criteria.

The best of all possible worlds

Of course, tradeoffs are inevitable. That old saw in engineering and software development says that you can't have faster, cheaper, and better—you can only have two out of three. "If you're trying to minimize the duration, you have to use overtime, and that means increasing your costs," says Khaled El-Rayes, an assistant professor in the Department of Civil Engineering at the University of Illinois at Urbana-Champaign. "If you're trying to improve quality, in many cases you have to pay more for that increase in quality."

How to reach a comfortable tradeoff between these conflicting objectives? That's the focus of the research that El-Rayes and his research assistant Amr Kandil are currently conducting, using NCSA machines to optimize the decision-making process. El-Rayes, who received an NSF Career Award for optimizing construction utilization of resources in transportation infrastructure systems, is developing an optimization model that can determine the optimal tradeoff between conflicting objectives. This is no simple problem. For each task involved in a large-scale construction project, there are at least three important criteria to consider—cost, duration, and quality. Plug in different combinations of possible values for each, and you can generate a large number of permutations involving different kinds of construction, equipment, and crews, the addition or omission of overtime, an off-peak work schedule, and other possible factors. With the average infrastructure project involving 600 or 700 different activities, the task of determining the optimal balance of duration, cost, and quality proves overwhelming for a human being.

Instead, El-Rayes uses a genetic algorithm-based model that allows him to generate a large number of possible construction resource utilization plans that provide a wide range of tradeoffs among project cost, duration, and quality and to eliminate the vast majority of suboptimal plans quickly. "At the end," he says, "what you want is a set of optimal tradeoffs which decision-makers can use to determine, according to their preferences, the best possible combination of resources."

Finding the optimal time-cost-quality balance for a construction project as a whole means considering several possible combinations for each individual construction activity. Because the number of construction activities for a large-scale project can number around 600 or 700, the pool of possible permutations, called the solution space, is extremely large.

This might mean, for example, that a longer project duration time is tolerable if cost or quality is a bigger concern, or that a reduced duration is a greater priority than cost or quality.

The advantage of this optimization model is its ability to transform the traditional two-dimensional time-cost tradeoff analysis to an advanced three-dimensional time-cost-quality tradeoff analysis. The introduction of the third dimension in construction projects is a challenging task, particularly when quality is itself a difficult factor to quantify. "The cost is simply dollar value, and so it is easy to aggregate by adding it all up," says El-Rayes. "Quality is more challenging."

El-Rayes' model, which incorporates quality, is currently based on data from the Illinois Department of Transportation (IDOT), which keeps records, for example, on the kind of utilized construction crews along with their measured performance in various quality metrics, such as compressive and flexural strength for concrete pavement work. Examining this data in aggregate, El-Rayes can determine how frequently and by how much a given combination of resources exceeded IDOT-specified quality limits, allowing him to assign a quality level to that specific construction crew and resource combination.

In the future, El-Rayes and his research team hope to be able to add even more factors for consideration, including safety, service disruption, and environmental impact. He would also like to make the process more user-friendly by including an interactive tool that would allow users to rank solutions based on weighting factors according to their preferences.

Optimizing the optimal

While El-Rayes' model, by automatically weeding out all less-than-optimum scenarios, makes the decision-making process easier for humans, there is no getting around the fact that it is still an enormous calculation. "If we had a project that included 700 activities, an average-sized construction process," explains El-Rayes, "and each activity had a potential three to five options each—and that's conservative—it would create a solution space which is exponential to the number of activities." It's a huge solution space, one which, El-Rayes estimates, would require around 430 hours of computation on a single processor. "Solving this problem wouldn't be feasible," El-Rayes says. "Nobody's going to wait 430 hours for the solution."

Using NCSA's Tungsten, El-Rayes and his research team, with the help of Nahil Sobh, who heads NCSA's Performance Engineering and Computational Methods group, are currently exploring how to parallelize his computations over a number of processors, so that rather than performing them on a single processor in a contractor's office or a state, local, or federal transportation department office, they can instead be distributed over a number of unutilized office processors, drastically reducing the run time to the duration of a weekend.

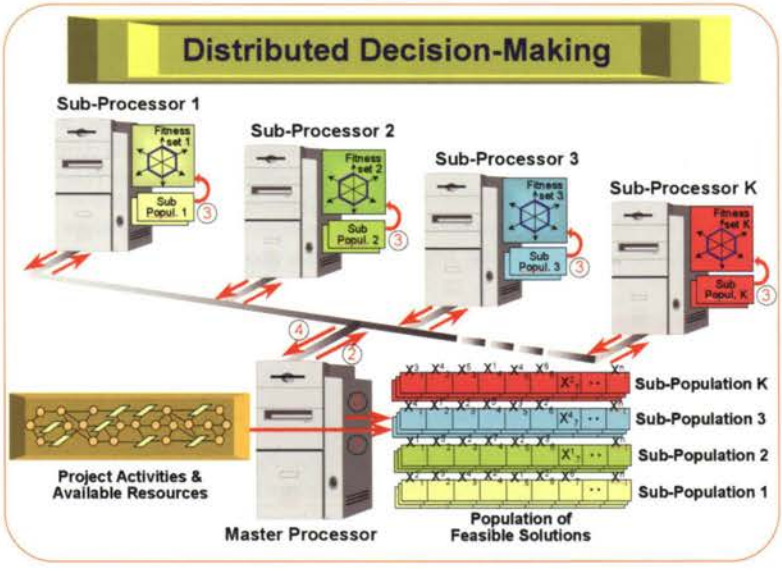
In his experiments on the NCSA Tungsten cluster, El-Rayes examined the required computational time for optimizing three construction projects of different sizes: 180 activities, 360 activities, and 720 activities, each of which he has analyzed on one processor and on multiple processors to a maximum of 50. So far, he says that parallelization has been successful in transforming the analysis of the largest project of 720 activities from an impractical problem requiring several weeks (430 hours) on a single computer to a feasible task that can be accomplished on a network of unutilized office computers over a weekend in 55 hours. "We don't even need 50 processors for this size project," he says. "For bigger projects, we might benefit from an increase in the number of processors, but the improvement starts to level off after maybe 10 to 15 processors, which is a reasonable number for an office to have available over a weekend." The problem he has chosen for these computations is a hypothetical highway construction project, but he says that the optimization model would be equally applicable to other kinds of large-scale projects, such as the construction of a convention center or a bridge, which would involve more different kinds of activities than would highway construction. What all large-scale projects have in common, however, is their complexity, and that is the problem El-Rayes hopes his computations will help solve.

"We want to transform an infeasible problem into a practical problem. That's what we're aiming for," says El-Rayes.

This research is supported by the National Science Foundation.

Access Online: http://access.ncsa.uiuc.edu/CoverStories/construction/
For further information: http://sftp.cee.uiuc.edu/people/elrayes/

Team members
Khaled El-Rayes
Amr Kandil
Nahil Sobh



A multi-objective genetic algorithm makes it possible to weigh more than two factors in determining the combinations of duration, cost, and quality that will produce the best possible outcome. Following the evolutionary biology model from which it takes its name, the algorithm strings together sets of values for each activity in a manner similar to that of a genetic code, with the eventual goal of producing the "fittest" result or the optimal combination of activities.

Nourishing new ideas

Story by J. William Bell
Illustrations by Blake Harvey

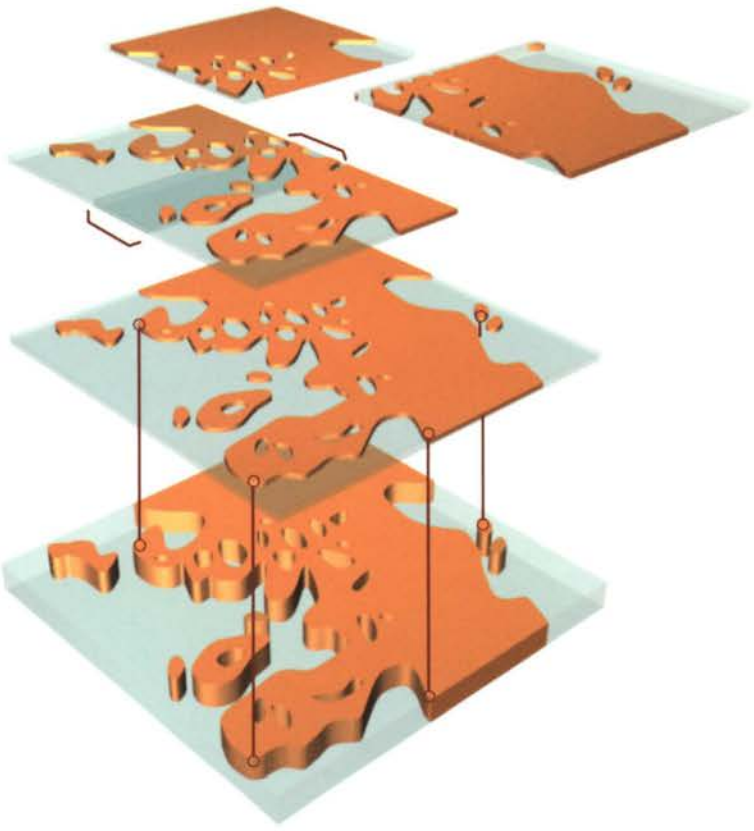
NCSA faces down myriad challenges in making sensor and instrument data useful.

Scientists are hungry for data because data nourishes their new ideas. The proliferation of sensor and instrument platforms brings new food to the table. Every day, researchers become less satisfied with the old process of taking a small statistical sample and extrapolating from there. They want to put hundreds, thousands, hundreds of thousands of sensors in the field and assimilate the data from all. They want to look at every piece of data that's ever been collected on the subject.

In other words, feeding a scientist often makes him more hungry, and that complicates matters tremendously. The data that scientists rely on often come from multiple instruments or sensor networks, and these data commonly vary in format or type. They are messy, poorly described, and difficult to integrate. Peter Bajcsy's team at NCSA investigates how to solve a variety of the problems that arise.

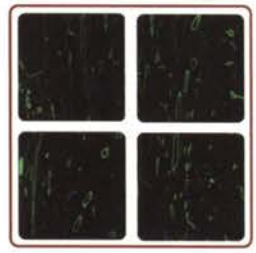
3D volume reconstruction

Data from medical instruments frequently capture images from slim cross-sections of neighboring parts of the same sample. To be useful, distortions have to be cleaned up. The images' registrations must also be aligned to ensure that matching points on different images of the same object represent matching points in the real world. Bajcsy's team is developing automated 3D volume reconstruction software that addresses these issues.



1 Slicing and imaging

Doctors cut tissue–samples of cancerous tumors or human organs, for example–into cross-sections that are only micrometers thick. They scan each cross-section several times with a specialized microscope. Each scan is three-dimensional, and neighboring scans slightly overlap one another.



2 Mosaicking

Multiple 3D volumes, acquired from the same cross-section, are stitched together.

3 Alignment

A registration algorithm compensates for any distortions that may have occurred during slicing and imaging. The algorithm aligns adjacent cross-sections that have been built from the mosaicked cross-sections.

4 Reconstruction

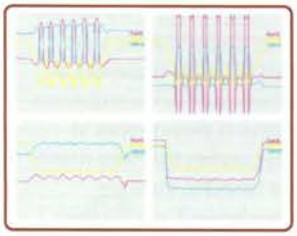
A 3D reconstruction is delivered. The team also offers tools to analyze volumetric density and shape, features that are frequently useful in diagnosis and medical research.

Gesture analysis

Controlling robots is just one possible use for gesture analysis. Bajcsy's team is working with the U.S. Navy and Chi Systems Inc. to devise an algorithm to control unmanned vehicles that are used on the decks of aircraft. They're also exploring the use of gesture-controlled robots to deploy sensor systems in hazardous environments like a nuclear reactor.

1 Output from body sensors

Sensors are placed on a person's arms. The sensors output their locations relative to one another on three axes. As a result, the data flowing from the sensors can be seen as building a series of angles and the order in which the angles occur.



2 Instruction translation and lexicon access

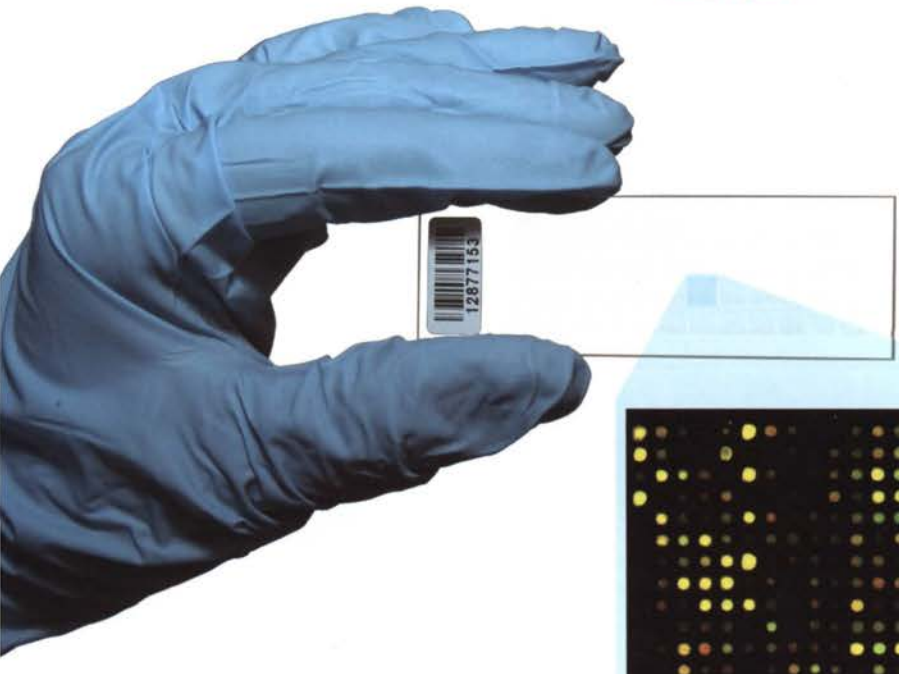
The received sensor data is processed by what the team calls a "gesture classifier." This algorithm compares the data's characteristics (the angles formed by the sensors) to a lexicon of about 20 gesture patterns. When the classifier recognizes a gesture pattern–say a right arm extended horizontally and a left arm repeatedly bending upward–it translates this match into a command.

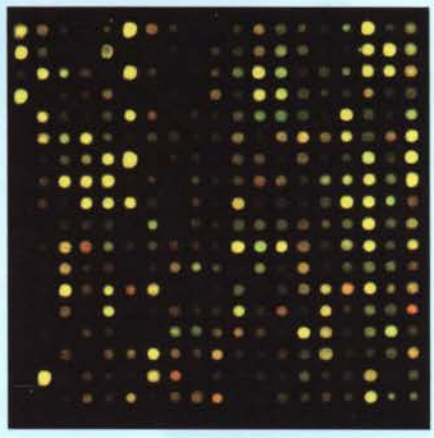
3 Robot action

The command is sent to the robot, where it is broken into a series of actions to be executed. For example, a "turn right" command might be converted to "adjust front axle 45 degrees to the right, spin all wheels 25 times, and then stop."

Microarray analysis

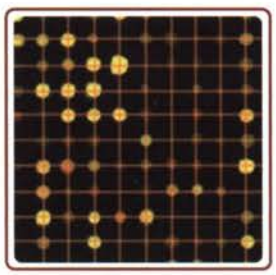
DNA microarrays, tiny chips covered in thousands of individual DNA strands, are crucial to contemporary genetic research. They allow researchers to identify which genes or sequences are in a particular DNA sample. Because a large number of strands is packed into a small amount of physical space, it's difficult to assess microarray data. A suite of algorithms by Bajcsy's team makes it easier.





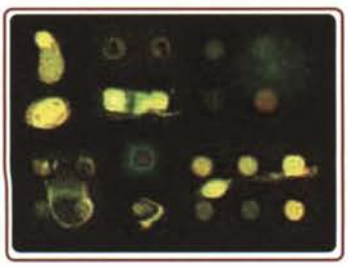
1 Preparing a sample

DNA is made of millions of bases, but those bases come in only four types. Each type has a complement, and those complements bind to one another. When a microarray is exposed to a sample, strands in the sample bind to strands on the microarray, called probes. Scientists know the sequences of the probes, so they can infer the sequences of strands that bind to those probes. In many microarrays, probes that have strands from the sample bound to them are designed to change color. Laser scanners detect these color changes and produce images for analysis.



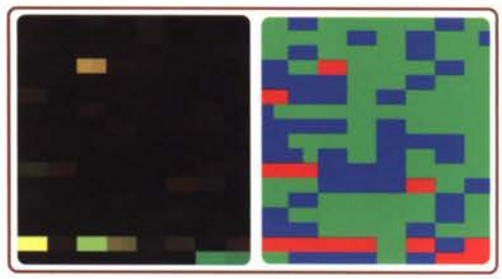
2 Grid alignment

These images are made up of rows and columns of colored spots. Spots are arranged in multiple grids. Due to fluctuations during preparation, these grids are inconsistently spaced and rotated. An algorithm locates and properly aligns each spot, despite this chaos.



3 Invalid spot detection

Spots are often misshapen. A spot might be shaped like a comet or a donut, or two spots might merge. Using multiple quality-control techniques and pre-defined quality thresholds, another software tool identifies invalid spots that do not meet specified criteria.



4 Segmentation and color extraction

Once valid spots have been identified, they are separated from the microarray image background. Spots are a combination of multiple colors, typically red and green. Each is processed pixel by pixel to determine the predominate color, statistically speaking. Colors can be further processed and clustered into like sets by the team's software, allowing for easier interpretation and extended analysis.

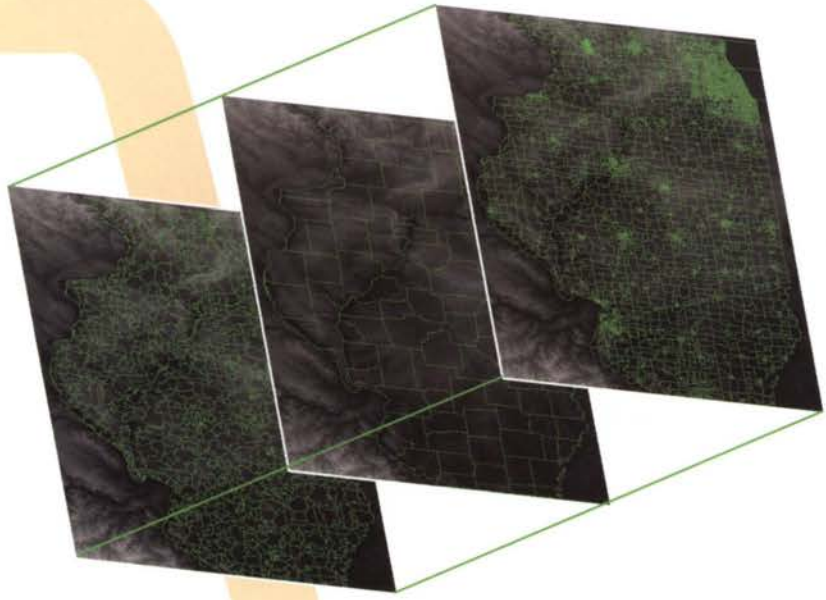
GIS data integration and decision support

Engineers, urban planners, and statisticians all use Geographical Information Systems (GIS) data to chart the physical and human features of the world. They might have maps and photographs called raster data, vector data that represent linear features like county boundaries or streams, and statistics in tabular form. The problem is that raster maps are made up of pixels, paths are described by vector data, and tabular data are entries in a spreadsheet. The maps often vary in scale, accuracy, and projection type, as well.

The team's I2K software allows scientists to integrate multiple maps, boundary data, and tabular data. It resolves the differences between various data types, performs analysis, and delivers information for GIS-based decisions.

1 Data integration

A geographical location is assigned to each point on the maps. Since a variety of data formats might be present, all of the maps are automatically converted to a single data representation based on an optimization of map parameters. For example, conversions are completed for varying resolutions, accuracy, and projection types. The maps are then mosaicked into a single image.



2 Boundary aggregation

The resulting image may be partitioned in any number of ways–by ZIP code, census block, or county, for example. These partitions each have statistical or geographic characteristics assigned to them. Partitions are aggregated into groups by the team's software, based on the attributes that they share. Perhaps all counties with a given number of robberies go together.

3 Evaluation and decision support

As aggregation occurs, deviations begin to arise. There were 17 robberies in a given ZIP code, but it's been grouped with ZIP codes that range from 15 to 30 robberies. To address this fact, users establish multiple error metrics. They might demand that the number of aggregations be maximized for easy interpretation, or they might demand the minimal amount of variance among ZIP codes within an aggregation. These metrics ensure that the optimal aggregation has been created.

Team members

Tyler Alumbaugh
Peter Bajcsy
David Clutter
Yi-Ting Chou
Peter Groves
Rob Kooper
Sang-Chul Lee
Young-Jin Lee
Sunayana Saha
David Scherba
Martin Urban

For further information:

http://alg.ncsa.uiuc.edu/do/documents/publications/
http://www.ncsa.uiuc.edu/Divisions/DMV/ALG/people/

This research is supported by NASA, the National Archives and Records Administration, the National Science Foundation, the National Institutes of Health, the University of Illinois College of Medicine, the Defense Advanced Research Projects Agency, the U.S. Navy, and the National Center for Advanced Secure Systems Research.

ACCESS serves as Election Protection HQ

On Election Day 2004 and the days leading up to it, NCSA's ACCESS center in Washington, D.C., served as headquarters for a nonpartisan voter-protection program. Election Protection 2004, created by a coalition of more than 100 civil rights and civic organizations, offered immediate advice and legal assistance to voters across the country.

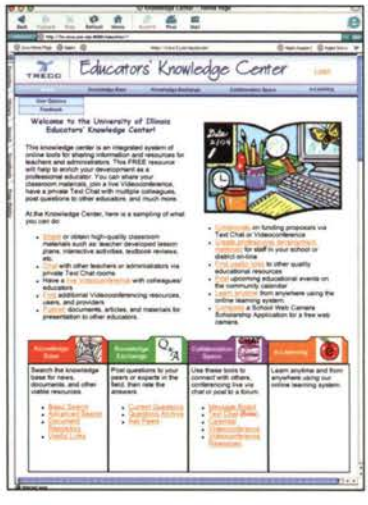
More than 300 volunteers—lawyers, law students, and paralegals—used ACCESS's collaborative communications offerings. Multiple teams worked from the center, setting up command and control operations, a press office, orientation and training for volunteers, call centers, data-entry facilities, large video displays, and reference libraries of states' election laws.

Election Protection 2004 was led by the Lawyers' Committee for Civil Rights Under the Law, the People for the American Way Foundation, and the National Coalition on Black Civic Participation.

Webcam giveaway boosts Educators' Knowledge Center

Since September 2004, the Technology Research, Education, and Commercialization Center (TRECC), which is administered by NCSA, has given about 100 Webcams to K-12 teachers and administrators throughout Illinois. The giveaway is meant to advance educators' technical capabilities, and the Webcams allow for easy, free videoconferencing. But the equipment and training educators receive are just the beginning.

"This is a new tool for many of these people, and we are a resource for them as they figure out 'Why do I need this?' and 'How do I use this?' says Nancy Komlanc, TRECC's director of education and training. "The school Web camera scholarship is also a way to jumpstart the University of Illinois' Educators' Knowledge Center."



The Knowledge Center is a free, integrated set of online tools for sharing information and resources. Among other things, educators can swap lesson plans, textbook reviews, and ideas for interactive activities. They can collaborate on proposals and projects via a text chat service. They can also—and here's where those Webcams come in—videoconference with colleagues in other buildings or other districts, using software and services provided by the Knowledge Center.

Numerous other communication tools are available in the Educators' Knowledge Center at: http://www.trecc.org/ed/. To apply for a Webcam, see: http://www.trecc.org/education/webcam/.

TRECC is funded by the Office of Naval Research.

NCSA technology powers start-up

Data mining and data analysis technology developed at NCSA are being brought to the marketplace by an Illinois-based start-up company, RiverGlass, Inc. The tools enable users to extract meaning from massive amounts of data of diverse types by searching for patterns, making predictions, identifying unusual features, optimizing complex problems, and visualizing the results. They are based on the D2K data mining software developed by NCSA's Automated Learning Group. Among the markets RiverGlass is targeting are law enforcement, homeland security, financial services, market intelligence, and network security.

RiverGlass
Real-Time Analytics

Dell joins Private Sector Partner program

Dell recently signed on as NCSA's newest Private Sector Partner. They join Boeing, Caterpillar, IBM, and Motorola, as well as Exxon Mobil, the program's most recent addition. Dell will leverage NCSA's expertise in several areas, including the development of monitoring tools and computer and network security.

In 2003, NCSA worked with Dell to install an Intel Xeon-based Linux cluster that employs more than 1,450 dual-processor Dell PowerEdge servers. The system debuted at number four on the Top500 list and currently is listed as the tenth-fastest supercomputer in the world. NCSA also completed a seven-teraflop cluster of 512 Dell PowerEdge 1850 servers.

With the addition of the new SGI Altix system, NCSA's machine room is now at capacity.

NCSA developing dependable grids

Grid computing is already widely used to facilitate advances in science and engineering. Its use in consumer services and critical infrastructure applications has been limited, however, because grids have not achieved the required dependability. Grids face threats from equipment or software failures, physical damage from natural disasters, and cyberattacks. Engineering a system to withstand all such attacks would be prohibitively expensive and complex.

NCSA, along with the University of Virginia, received a $2-million NSF Information Technology Research grant to counter these problems and develop dependable grid computing technologies. The NCSA/Virginia team aims to create a survivable system, one that provides one or more alternate services in the event an attack or failure disrupts the primary service.

NCSA's Jim Basney will lead the center's participation in the project, including establishing a "grid dependability lab" for evaluating software as well as developing dependable software components based on the Globus Toolkit. Testing and experimentation will span the test labs at Virginia and NCSA, and the project will develop software based on multiple grid software toolkits.

Five In Focus

NCSA develops the tools and technologies that safeguard the data, computers, and applications of cyberinfrastructure. Following are five ongoing NCSA security research projects:

1. **Mining Alarming Incidents in Data Streams (MAIDS):** *led by Michael Welge, head of NCSA's Automated Learning Group.* The goal of MAIDS is to use data mining to monitor trends and characteristics in dynamically streaming data—such as data on electrical power loads, network traffic, and from environment monitoring—and to detect intrusions into the data stream.

2. **Multicast Security and Survivability (MulticastSS):** *led by William Yurcik; Jun Wang, technical project manager.* MulticastSS is developing strategies and prototype software for user-configurable group communications that can survive different failure scenarios, including physical and virtual attacks.

3. **Cluster Security (Cluster-Sec):** *led by William Yurcik; Forrest Xin Meng, lead developer.* The Cluster-Sec team has developed the world's first cluster security monitor, NVisionCC, which provides a single-screen overview of activity on increasingly large high-performance computing clusters.

4. **Security Incident Fusion Tools (SIFT):** *led by William Yurcik; Adam Slagell, technical project manager.* The goal of the SIFT project is to help security engineers determine when a network is under attack, what is being attacked, and what form the attack is taking.

5. **Secure Email List Service (SELS):** *led by Himanshu Khurana.* The SELS protocol provides confidentiality, integrity, authentication, and anti-spamming tools for email list services.

3. A vision for cluster security

Clusters of computers are capable of performing trillions of calculations each second, storing trillions of bytes of data, and communicating over high-speed networks. Unfortunately, the power of clusters makes them an attractive target for hackers. As clusters grow to include thousands of nodes, security monitoring becomes extremely complex.

NCSA's Cluster Security (Cluster-Sec) team is developing the first software tool, NVisionCC, specifically designed to monitor cluster security. NVisionCC (CC stands for cluster computing.) collects and synthesizes data from heterogeneous sources, presents the information visually, and alerts users to potential security problems.

The Cluster-Sec team realized that the many nodes on a cluster fall into a small number of classes. Most of the nodes are allocated to run jobs (compute nodes), some are used to access the cluster, compile software, and submit and monitor jobs (head nodes), and some are storage nodes. Profiles can be set for each class, defining the parameters of acceptable, secure use, including the processes that are allowable on that type of node and the ports that can be used. NVisionCC compares data on each node to the acceptable profile for that node's class. Any activity that falls outside the defined profile is flagged for a system administrator to examine.

William Yurcik, seated at left, and the Cluster-Sec team.

Passive tracer in a stratified jet

Motions in stably stratified environments—such as oceans, lakes, and some parts of the atmosphere—are often highly anisotropic. In other words, their properties vary depending upon the direction from which the properties are measured. This feature is due to the presence of buoyancy forces that act in only one dimension. In laboratory experiments, for example, the fluid flow in the wake of a towed sphere evolves toward a highly stable configuration of vortices with nearly horizontal motions called pancake vortices.

Analytical theory and preliminary numerical simulations—like those shown here illustrating the concentration of a passive tracer in a stratified jet—imply something more complicated. They suggest that this stable state represents a dynamic balance between inertial and viscous forces. Therefore, the stable state may be a manifestation of the relatively low Reynolds numbers attainable in the laboratory.

Using a pseudo-spectral numerical model on a 1,024-by-512-by-256 grid, a cross-country team of researchers recently achieved high enough Reynolds numbers to resolve secondary instabilities in an idealized jet flow. The team includes Kraig Winters of the Scripps Institution of Oceanography and the University of California at San Diego, James Riley of the University of Washington, and Steve de Bruyn Kops of the University of Massachusetts at Amherst. Calculations were completed on the Scripps Institution's COMPAS (Center for Observation, Modeling, and Prediction at Scripps) cluster. Visualizations were created by David Semeraro, assistant director of NCSA's Visualization and Virtual Environments group.

The secondary instabilities appear to disrupt the sequence of events leading to quasi-steady pancake motions. At the Reynolds numbers the team is able to simulate, however, they do not radically alter the final character of the slowly varying flow. The secondary instabilities are shear-driven Kelvin-Helmholtz instabilities. They appear as small patches of alternating light and dark lines in the image.

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
605 East Springfield Avenue
Champaign, IL 61820-5518

http://www.ncsa.uiuc.edu